THE NATURE OF INFORMATION PROCESSING IN SPEECH PERCEPTION[1]

U. Thein-Tun* and D. Burnham**

*School of Communication Disorders
La Trobe University

**School of Psychology
University of New South Wales

ABSTRACT - Since several pairs of non-speech sounds on each end of an acoustic continuum can be perceived categorically in the same way as a minimal pair of two phonemes on each end of a speech continuum are, the notion of speech perception has been debated either as a special or non-special mode of information processing. On the strength of phonetic cue trading as an experimental tool, it can be shown that the higher the linguistic level is from words to sentences, the more special is the mode of information processing in speech perception.

INTRODUCTION

Every phonological contrast is cued by several distinct acoustic properties of the speech signal. Within limits set by the relative perceptual weights and by the effective ranges of these cues, a change in the setting of one cue (which by itself would have led to a change in phonetic percept) can be offset by an opposite or complementary change in the setting of another cue so as to maintain the original phonetic percept. This phenomenon is known as phonetic cue trading (Fitch, Howes, Erickson & Liberman, 1980).

It has been repeatedly shown by the realists (those in the speech-is-not-special camp) that speech perception cannot be regarded as a special mode of information processing on the basis of categorical discrimination and identification alone (e.g. Jusczyk, Pisoni, Walley & Murray,1980; Dooling, Okanoya & Brown, 1989; Dooling & Brown, 1990) nor on the basis of a special innate mode of processing in human infants (Aslin, Pisoni, Hennessy & Percy, 1981; Jusczyk, Pisoni, Walley & Murray, 1980; Burnham, 1986). Since it has been established that speech cannot be considered special on the basis of evidence from categorical perception, the mentalists (those in the speech-is-special camp) have been arguing for their case on the basis of a different phenomenon, namely, cue trading relation. They claim that every phonetic contrast is cued by many different acoustic properties and the integration or trading relation of these constituent cues by the listener in perceiving speech sounds is unique to speech, even though categorical identification may not be.

Best, Morrongiello and Robson (1981) investigated the trading relation between silent closure duration and F1 transition onset frequency as cues to the existence of the stop in the "say" - "stay" contrast. They constructed sinewave analogs of the "say" - "stay" continuum by presenting a noise resembling [s]-frication followed by varying periods of silence and a sinewave portion whose component tones imitated the first three formants of the periodic portion of the speech stimuli. The sinewave stimuli were presented to listeners in an AXB discrimination task. Some of the subjects were told that the stimuli were intended to sound like "say" or "stay" whereas others were only told that the stimuli were computer sounds. The results showed that those who were instructed to treat the stimuli as speech sounds demonstrated a trading relation between silence and F1 onset frequency while those who were instructed to treat the stimuli as non-speech sounds did not demonstrate any perceptual pattern that resembled a phonetic cue trading relation. These results suggest that the speech mode of processing is a special mode because it incorporates cue trading while other more auditory modes of processing do not invoke the special use of cue trading relations. Since the publication of Best et al. (1981) the mentalists have been continually reporting a flurry of studies on cue trading using the Best et al. dual processing technique (e.g. Repp, 1981a & b; Parker, Diehl & Kluender, 1986; Kluender, Diehl & Wright, 1985; van Heuven, 1987). Although their results were not overly convincing and showed a certain degree of cue trading in the non-speech mode similar to that found by Thein-Tun (1986 & 1987a & b), they nevertheless made their judgement in favour of the speech-is-special view. The use of the dual perception method (i.e. comparing the perception of the same stimuli in speech and non-speech modes by two different groups) in their studies seems to be a very desirable approach. Nevertheless, the following remarks on these studies as well as the majority of psychophysical studies conducted

during the last three decades or so seem to be timely:
(i) The phonological contrasts investigated are in the domain of either stop consonant voicing contrasts (e.g. Parker, Diehl & Kluender,1986; and Kluender Diehl & Wright, 1985) or fricative affricate contrasts (e.g. van Heuven, 1987; Howell and Rosen, 1987)) which are highly temporal in nature , i.e., for those consonants, silent gap or VOT rather than a spectral property, happens to be the main cue. (This criticism in fact is so applicable that one could be forgiven for concluding that stop consonant voicing contrasts and silent gap (fricative-affricate) contrasts are the only phonological contrasts in English and other languages of the world).

(ii) When the contrast depends mainly on a temporal cue such as a silent gap or VOT, the occurrence of cue trading relation in the non-speech mode is affected not only by the constituent cues underlying the phoneme contrast but also by other higher level cues such as speaking rate of the phrase (e.g. Howell and Rosen, 1987). Thus in such cases a meaningful interpretation of the results becomes remote.

(iii) Possibly due to the purely psychophysical impetus of these studies, the majority of the investigators have failed to include relevant psycholinguistic factors in their designs. And unfortunately they have confined their investigations to isolated words or syllables and have tended to ignore the fact that speakers do not normally produce or hear isolated words or syllables in their everyday language experience. One other unfortunate methodological aspect in most of these studies is the heavy reliance on the discrimination task (either in an AX or ABX or AXB paradigm) which is an experimental tool with no real correlate in everyday language experience.

In view of these remarks the following research speculations are not only in order but also provide a very interesting and tantalising challenge:

(a) The existence of cue trading in both speech and non-speech modes may depend on the type of particular phonological contrast investigated; unlike the temporally oriented stop voicing contrast, the spectrally based place or manner contrast is more speech specific and its parallel cue trading situation may not exist at all in the non-speech mode.

(b) the extent of cue trading for particular contrasts may also depend on the phonetic environment, i.e., the linguistic level at which the contrasts occur, and the linguistic maturity of the listener.

These speculations are based on the general assumption that whether a cue trading relation for a phonological contrast can be found in both phonetic speech mode and its corresponding auditory (non-speech) mode depends primarily on the type of phonological contrast and secondly on many other experimental factors such as phonetic environment and linguistic (syntactic) levels at which the stimuli occur. In this regard, it seems highly likely that stop voicing contrasts are more temporally and acoustically oriented whereas manner and place contrasts are more spectrally and system oriented. Therefore it would be expected that under comparable conditions cue trading should occur in both speech and non-speech modes for voicing contrasts but only in the speech mode for place and manner contrasts.This project was designed to investigate these research speculations using the multilinguistic level method devised by Thein-Tun (1986, 1987a & b). However, rather than investigating the voicing contrast (Thein-Tun, 1986 & 1987), we investigated the set of phonetic parameters that constitute both place and manner contrasts for identifying /p/ and /h/ in the "pen" - "hen" continuum. This contrast has the advantage in that it does not need articulatory stoppage or VOT as an acoustic cue.

The arguments which centre around the cue trading phenomenon in this regard are reminiscent of the debates that used to revolve around the categorical perception phenomenon. The mentalists and realists have been playing a tennis game (Massaro, 1987; p. 46) on whether speech perception is special or not with the phenomenon of categorical perception as the net. At the present stage workers in both the camps seem to have come to some compromised understanding that categorical perception is dependent upon so many experimental (e.g., Repp,1984 & 1987), psycholinguistic and psychoacoustic (e.g., Massaro, 1987; Best, Studdert-Kennedy, Manuel & Rubin- Pitz, 1989; Miller &Volaitis, 1989) factors that it is neither a fundamental characteristic of speech nor is it valid to argue for or against the special speech mode on its basis. In order to investigate that research speculation, we believe psychophysicists will have to (i)explore other phonological contrasts (than stop voicing and stop gap fricative-affricate contrasts) such as place of articulation and manner of articulation contrasts which have not been fully investigated in the cue trading context yet, and (ii) incorporate in the research design the relevant linguistic and psycholinguistic factors which are part of people's everyday speech perception experience.

METHOD

Construction of the "pen - hen" continuum

Three 8-step "pen - hen" continua were constructed first. The phonetic parameters that change proportionally along eight steps were initial Fo pattern (from the high falling initial [p] position to the level [h] position), spectral energy spread of the initial aspiration phase (from the 1800Hz - 8000Hz region to the level 1800Hz - 5000Hz region) and initial F1,F2 and F3 transitions (from the rising bilabial [p] position to the level neutral [h] position). In order to maintain the contrastive cue parameters along the continua and exclude the silent gap temporal cue, the [p] release burst was not included in the synthesis. The reason for this exclusion was to make the basis of the continuum as spectral rather than temporal as possible. (The rationale for this is that spectral cues are more speech oriented than temporal cues). The schematised spectrographic patterns for the formulation of the "pen-hen" continuum together with the average values (of component durations, formant trajectories, Fo patterns, aperiodic frequency regions) taken from a sample of five male Australian native speakers and used as basic parameters in our experiments are given in Figure 1.

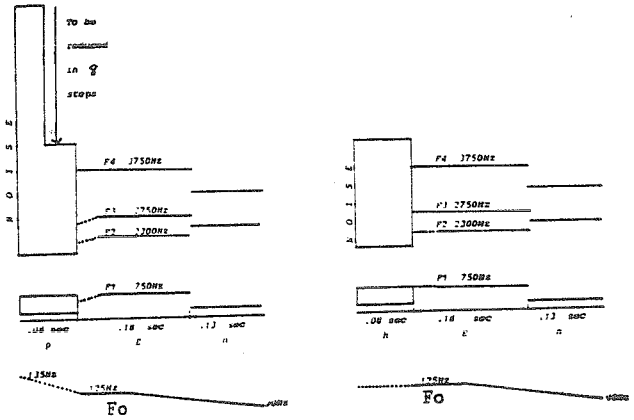

Figure 1. Schematised spectrographic patterns of "pen" & "hen"

The synthesis software used was CSRE (Canadian Speech Research Environment) version 4.1 which utilises the well-known Klatt synthesis system. The resultant synthesis quality of the continua was so good that the listeners could hear and identify the end points as intended in the synthesis with no anchoring. The intensity of the aspiration phase and that of the following [ɛn] portion of the basic continuum with the best [pɛn] and the best [hɛn] on each end of the continuum was given the nominal value of 0dB. Out of the basic eight step continuum thus created, further two continua were created by changing the relative intensity of the aspiration phase and the following [ɛn] phase by 4dB and combining the two phases in three ways:

	Aspiration phase amplitude		[ɛn] phase amplitude
	Aperiodic or A phase		Periodic or P phase
[h] biased	+4dB	combined	-4dB
Non-biased	0dB		0dB
[p] biased	-4dB		+4dB

The total of 24 steps from these three continua served the word level identification test.

Auditory, Phonetic and Sentence Level Stimulus Construction

The above 24 word level basic stimuli were transformed into white noise plus sinewave analogs by reducing the bandwidth of all the formants to near zero level (as low as the system can accept). This task was in fact accomplished by increasing the intensity of the voice source as a parameter. These 24 steps of the three analog continua served as 24 independent test stimuli both at the phonetic level and the auditory level. The phonetic mode of processing was instigated by telling the subjects that the stimuli were speech sounds "pen" and "hen" transformed into whistling. The

auditory mode of processing was accommodated by telling the subjects that the stimuli were non-speech sounds (sound 1 and sound 2). For the sentence level speech mode, the original 24 stimuli were carried by short synthesised sentences. The synthesised carrier sentences were within six to nine syllables with no /p/ or /h/ in syllable initial position and no semantic cues to the /p/ - /h/ decision or perception. The target stimuli appeared as the last word of the sentence, e.g., "Mary would not like the ---- " and "They may see about the ----" and so on. The time between the target stimulus word and the preceding word was kept constant for all test sentences. Since synthesised target words would not be entirely compatible with natural speech, carrier sentences had to be synthesised.

Test Procedure and Subjects

The stimuli were presented to subjects on headphones from an audio tape recorder. The tape recorder was controlled by an IBM-compatible computer programmed to play, pause and stop the tape as necessary and to record subjects' responses. Each subject was tested individually on only one linguistic level (auditory, phonetic, word or sentence). Twenty adults and 20 children were tested at each level. All were native Australian (English) speakers with no hearing or neurological impairment. For each level the 3 X 8 = 24 stimuli were arranged in four blocks on the audio tape in different random orders in each block. Thus each subject heard and responded to each stimulus four times.

On a response panel in front of them was a "ready" key and above it side-by-side were two response keys, one with a drawing of a pen and one with a drawing of a hen. For the auditory level test, the pictures were replaced by colour-coded keys, e.g., one red (for sound one) and one blue (for sound 2). The rationale for using pictures and colour-coding is to avoid orthographic interference, the degree of which may well differ between adults and children.

On any particular trial the ready key lit up to indicate that the computer was ready. Subjects then pressed the "ready" key to indicate their attentiveness. The trial was then presented and the two response keys lit up at the onset of the critical stimulus to indicate that the subject should respond as quickly as possible. The lights remained on for two seconds within which time the subjects had to make their response. (If the subject was too slow all three lights flashed to encourage quicker responding on subsequent trials). If the ready key was not pressed again within one second of the two second period, the audio tape would pause automatically, waiting until the subject was ready. In this way subjects could rest at will and attentiveness on each trial was maximised. The children were between six and seven years of age . Testing of children was identical to the above except that the experimenter had a button press input to the computer. This had to be pressed after a subject had pressed the ready key as an extra precaution against inattention. If necessary children were given breaks between the blocks. Before testing each subject was given two pre-tests. Firstly, puretone tests were given in order to screen out those with hearing impairment. Secondly, subjects in all four level groups were given a trial identification test with the ten steps from the basic word level continuum (0dB in both aspiration and /ɛn/ phases). Subjects must have had responses with an expected category boundary in order to participate in the experiment. In addition, this pre-test ensured that subjects learned the trial sequence and were taught to make responses within the two second response period.

RESULTS AND SUMMARY DISCUSSION

Two measures were taken, one of the categoricality of subject responses (Burnham, Earnshaw & Clark, 1991) and one of the strength of cue trading (Thein-Tun, 1987 b). The results for the categorical scores are shown in Figure 2. As can be seen adults' scores were generally more categorical in the non-biased condition than the other two conditions. Thus the p-biased and h-biased conditions subjects were less definite in the categorical identification of the sounds. More importantly it can be seen that as the linguistic level increased the degree of categorical perception decreased. Children's perception was generally less categorical and the above effects were less apparent. These observations are borne out by a 2 (age) X 4 (linguistic level) X 3 (No/p/h bias) analysis of variance: sentence level stimuli were perceived less categorically than word level stimuli, $F(1,147) = 8.72$, and sentence and word level less categorical than phonetic and non-linguistic sinewave level, $F(1,147) = 20.07$, especially for the adults, $F(1,147) = 7.58$. The results for the cue trading scores are shown in Figure 3. As can be seen there was much more cue-trading for the h-bias condition than for the p-bias and this was true for both the adults and children. Moreover, for both adults and children the degree of cue trading for h-bias was greater for sentence and word levels than phonetic and sinewave levels, $F(1,147) = 16.99$, but, contrary to expectations, h-bias cue

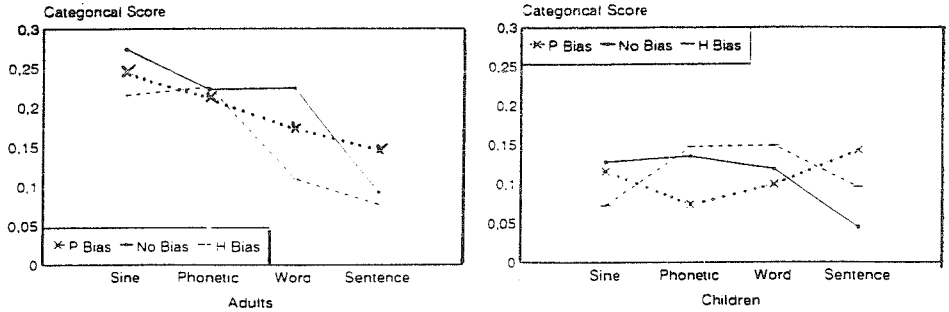

Figure 2. Pen-Hen experiment - categorical scores

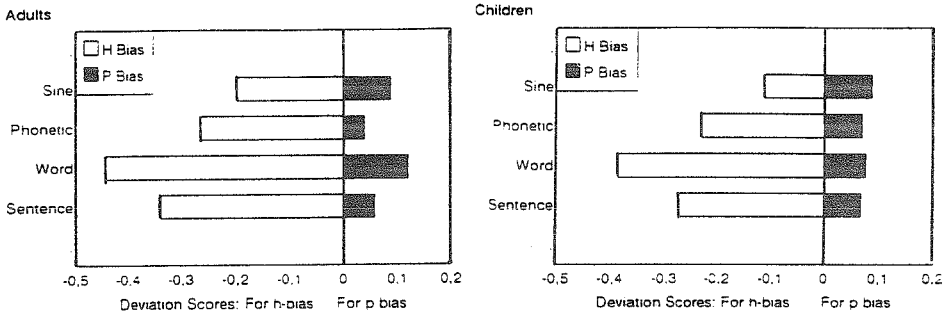

Figure 3. Pen-Hen experiments - cue trading scores

trading at the sentence level was less than at the word level, $F(1,147) = 4.17$. In the case of the h-biased continuum, the results are in agreement with Thein-Tun's (1987a & b) finding that the higher the linguistic level, the stronger the cue trading and the weaker (or non-existent) the categorical identification. The greater degree of cue trading for h-bias for sentence and word levels also indicates that the nature of cue trading is a speech specific characteristic. In view of this remark, two queries remain; why was the degree of cue trading for the sentence level not greater than that for the word level and why did the degree of cue trading for p-bias remain the same or behave at random for all the linguistic levels. The first query can be explained by the fact that carrier sentences were not included in the pre-test trials and hence the subjects' concentration on the target words was obviously impaired by unfamiliar sentences. Regarding the second query, it is rather obvious that the intensity manipulation of -4dB aperiodic phase and +4dB periodic phase for the p-biased condition was not acoustically potent enough to invoke delayed boundary changes from "p" to "h" although the reverse arrangement worked for the h-biased condition. This explanation remains to be ascertained by further experiments.

NOTE

(1) This project was funded by the Australian Research Council.

REFERENCES

Aslin, R.N., Pisoni, D.B., Hennessy,B.L., & Percy A.J. (1981) Discrimination of voice onset time by human infants: New findings and implications for the effects of early experience. Child Development, 1981, 52, 1135 - 1145.

Best, C.T., Morrongiello, B., & Robson R. (1981) Perceptual equivalence of acoustic cues in speech and non-speech perception. Percept. & Psychophys. 29, 191 - 221.

Best, C.T., Studdert-Kennedy, M., Manuel, S., & Rubin-Pitz, J. (1989) Discovering phonetic coherence in acoustic patterns. Percept. & Psychophys. 45, 237 - 250.

Burnham, D.K., (1986) Developmental loss of speech perception: Exposure to and experience with a first language. Applied Psycholinguistics, 1986, 7, 207 - 239.

Burnham, D.K., Earnshaw, L.J., & Clark, J.E. (1991) Development of categorical identification of native and non-native bilabial stops: infants, children and adults. J. Child Lang. 18, 231-260.

Dooling, R.J., Okanoya, K., & Brown, S.D. (1989) Speech perception by budgerigars (Melopsittacus undulatus): The voiced voiceless distinction. Percept. & Psychophys.46, 65 - 71.

Dooling, R.J., & Brown, S.D. (1990) Speech perception by budgerigars (Melopsittacus undulatus): Spoken vowels. Percept. & Psychophys. 47, 568 - 574.

Fitch, H.L., Halwes, T., Erickson, D.M., & Liberman , A.M. (1980) Perceptual equivalence of two acoustic cues for stop consonant manner. Percept. & Psychophys. 27, 343-350.

Howell, P., & Rosen, S. (1987) Perceptual integration of rise time and silence in affricate/fricative and pluck/bow continua. In Schouten, 1987, 173 - 180.

Jusczyk. P.W., Pisoni, D.B., Walley, A., & Murray, J. (1980) Discrimination of relative onset time of two-component tones by infants. J. Acoust. Soc. Am. 67, 262 - 270.

Kluender, K.R., Diehl, R.L. & Wright, B.A. (1985) Perception of duration of medial silent intervals in speech and nonspeech signals. J. Acoust. Soc. Am. 77, S27 (Abstract)

Massaro, D.W. (1987) Psychophysics versus specialized processes in speech perception. In Schouten, 1987, 46 - 65.

Miller, J.L., & Volaitis, L.E. (1989) Effect of speaking rate on the perceptual structure of phonetic category. Percept. & Psychophys. 46, 505 - 512.

Parker, E.M., Diehl, R.L., & Kluender, K.R. (1986) Trading relations in speech and nonspeech. Percept. & Psychophys. 39, 129 - 142.

Repp, B.H. (1981a) Auditory and phonetic trading relations between acoustic cues in speech perception: Preliminary results. Haskins Laboratories Status Report on Speech Research, 1981, SR-67/68, 165 - 189.

Repp, B.H. (1981b) Phonetic trading relations and context effects: New experimental evidence for a speech model of perception. Haskins Laboratories Status Report on Speech Research, 1981, SR-67/68, 1 - 40.

Repp, B.H. (1984) Categorical perception: Issues, methods, findings. In N.J. Lass (Edt.), Speech and Language: Advances in Research and practice, Vol. 10, New York: Academic, 1984, 243 - 335.

Repp, B.H. (1987) The role of psychophysics in understanding speech perception. In Schouten, 1987, 3 - 27.

Schouten, M.E.H. (Edt.), (1987) The Psychophysics of Speech Perception, NATO ASI Series D: Behavioural Sciences - No. 39. Dordrecht: Martinus Nihoff, 1987.

Thein-Tun, U. (1986) Multi-Linguistic Level Cue-trading Relations for Initial Stop Voicing by Normal and Hearing Impaired Listeners. Ph. D. Thesis, Macquarie University ,1986.

Thein-Tun, U. (1987a) Cue-trading Relations for Initial Stop Voicing Contrast at different Linguistic Levels. The Proc. 11th Inter. Cong. Phon. Sc, Tallinn, Estonia, U.S.S.R., 1987, Vol. 5, 354 - 357.

Thein-Tun, U. (1987b) Phonetic Cue Trading, Categorical Perception and the Order of Speech Processing. Speech Communication, 1987, 6, 353 - 362.
van Heuven, V.J. (1987) Reversal of rise time cue in the affricate/fricative contrast: an experiment on the silent sound. In Schouten,1987, 181 - 187.